

# 本周周报

解聪

2013.11.18-2013.11.24

## 本周工作

### 一、数值属性分段的方法

将人群（企业）按照某种属性  $P$  分为  $K$  段，使得各段用户（企业）行为分布的熵总和最小。  
Entropy minimize 可以用来进行聚类或者模式识别，可参考：Pattern discovery via entropy minimization 以及 An Unsupervised Clustering Method using the Entropy Minimization 等文章。

#### 方法一：

如果就要将数据分成 10 份，是不是可以先将数据分为 2 份，在两份的基础上分 3 份.. 依次类推。

每次划分，在上一步的基础上选取已有的分类中熵最大的那一类，对该类中进一步划分。

问题：该方法不能够保证取到全局最优的结果。

#### 方法二：参考利用最小熵进行聚类的方法：

1. 将数据随机分为  $K$  类
2. 重新分配点  $p$ ，看  $p$  在哪一类中可以使得最小，将  $p$  调整到该类中。
3. 迭代步骤二，当每个类熵的变化小于某一阈值时，停止迭代。

由于数值属性是连续的，调整边界的时候计算量较大。为了简单处理，将数值属性变为离散的数据。比如现将数据按照属性划分为  $N=1000$  份，在对这  $N=1000$  份进行类似于聚类的操作。

划分的问题不同于聚类的地方在于，划分后每一类中的属性要是值是相邻的。

因此算法如下：

1. 将  $N$  个数据单元按照相邻的属性值随机分为  $K$  类。
2. 调整每一类的边界的一个数据单元，使得调整后的其相邻两类的熵总和最小。
3. 迭代步骤 2。

**实现：**首先将数据按照数值属性，比如利润，在 -50000~+200000 之间均匀划分为 100 个的数据单元分别记为  $x_1, x_2, \dots, x_{100}$ 。

初始划分为 10 类：  $x_1-x_{10}$ ,  $x_{11}-x_{20}$ ,  $x_{21}-x_{30}$ , ... ,  $x_{91}-x_{100}$ 。每一类都计算其在 527 个行业中的分布情况并计算每类的熵的大小：

4.583717252 , 5.160341708 , 5.32968533 , 5.475500835 , 5.471911388 , 5.463582542 ,  
5.422399728, 5.432450486, 5.079514452, 4.671244554

10 类熵的总和为： 52.09035

第一次迭代后边界点：  $x_{10}$ ,  $x_{19}$ ,  $x_{30}$ ,  $x_{41}$ ,  $x_{51}$ ,  $x_{61}$ ,  $x_{71}$ ,  $x_{81}$ ,  $x_{90}$ ,  $x_{100}$ 。

第二次：  $x_0$ ,  $x_{10}$ ,  $x_{18}$ ,  $x_{29}$ ,  $x_{42}$ ,  $x_{52}$ ,  $x_{62}$ ,  $x_{72}$ ,  $x_{82}$ ,  $x_{90}$ ,  $x_{100}$

....

20 次: x0, x11, x12, x21, x43, x58, x59, x78, x88, x89, x100

20 次后每类熵的大小为:

4.60663769, 4.223543845, 5.218784071, 5.482554176, 5.472277005, 5.081148315,  
5.471266345, 5.146281286, 3.75263629, 4.733618762

10 类熵的总和为: 49.18875

问题: 该方法调整边界元素的时候, 现在还不能保证取得全局最优。另外收敛可能太慢。

### 方法三:

对上述方法进行改进,

即第二步调整数据单元时, 不仅仅考虑边界上的数据单元。穷举所有可能两段之间的边界情况, 求得全局最优。

问题: 该方法的问题是计算量太大。

## 二、对不同属性作为人群划分依据的比较

不同属性对数据会产生不同的效果, 我们要找出最具代表性的属性对数据进行分类。

比如按照年龄对淘宝数据划分后, 得到男女在购买类目上的分布的熵为: 男 M: 30.04  
和女 F: 28.04

12 星座划分后熵为: 24.29, 24.75, 24.03, 24.14, 24.91, 24.17, 25.58, 22.1, 26.01,  
25.74, 26.61, 24.54

之前认为某属性划分后, 人群的熵越小, 说明该属性能够越准确地进行人群分类。

可以发现性别划分得到的熵的平均值比星座划分的熵的值大。是不是说星座划分得到的人群混乱程度更小?

我个人感觉不能这么理解。因为注意到性别将数据分为了两份, 每一份占数据的 1/2, 星座的划分为 12 份, 人数平均是 1/12。因为第一种方法每类中的人数多, 自然混乱程度就增加了。但是从之前可视化视图中看, 性别的确是划分人群的一个有效的属性。因此不能仅仅看熵的大小。

因此需要定义不同属性划分人群的有效性。参考的因素有以下几点:

1. 划分的人群数。
2. 划分后每个人群熵的大小。
3. 划分后的人群之间熵的差异, 可以用相对熵来衡量。比如  $D = F \cdot \ln(F/M)$

我们需要设计一个标准整合以上各种因素, 来考虑一个属性对于人群划分的效果。具体怎么做还在考虑中。

### 下周工作:

继续解决本周的两个问题。

研究不同属性之间的互信息对人群分类的影响。